JORNADAS SOBRE BIBLIOTECAS NACIONALES

"LAS BIBLIOTECAS NACIONALES DEL SIGLO XXI"

Biblioteca Valenciana, 18, 19, 20 y 21 de mayo de 2005

PONENCIA

TÍTULO: RECURSOS DIGITALES: UN RETO PARA LAS BIBLIOTECAS NACIONALES.

AUTOR: XAVIER AGENJO (Fundación Ignacio Larramendi).*

1. Introducción

Son muy numerosos los artículos que en España se están publicando acerca de lo que suponen y lo que pueden suponer los recursos digitales para el mejor funcionamiento de todo tipo de bibliotecas. Sin embargo, la mayoría de ellos, por no decir la práctica totalidad, se refieren a los recursos digitales como algo que la biblioteca *consume* y pone a disposición de sus usuarios, y rara vez, como algo que la biblioteca *produce*[1], también para ponerlo a disposición de los usuarios. Ya hace años, en un texto que no he podido actualizar hasta el

* xavier.agenjo@larramendi.es

[1] En este sentido puede leerse, al menos como el análisis de un '*case study*' el artículo de Francisca Hernández, Carlos Wert, Ignacio Recio, Begoña Aguilera, Walter Koch, Martin Bogensperger, Peter Linde, Georg Günter, Bob Mulrenin, Xavier Agenjo, Robin Yeats, Luciana Bordoni, Fabrizio Poggi: *XML for Libraries, Archives, and Museums: The Project Covax.*//*Applied Artificial Intelligence* .-17 (8-9): 797-816 (2003).

Este artículo describe una completa metodología, realizada por europeos y dentro del programa IST acerca de cómo es perfectamente posible analizar y diseñar las soluciones técnicas precisas para permitir el acceso, a través de Internet, a descripciones de documentos de archivos, bibliotecas y museos codificadas de forma homogénea, basándose en la aplicación de SGML/XML, y crear bases de datos, utilizando tecnología avanzada. El proyecto COVAX (*Contemporary Virtual Archives in XML*) se inició en 2000 y concluyó en 2002. Puede consultarse aún su sitio en la red: www.covax.org

momento[2], hacía yo referencia a esta situación, y veo que mis predicciones se han ajustado, por desgracia, a la realidad.

No hace mucho tiempo que, desde determinada Secretaría de Estado, se criticaba la tendencia, tanto del sector público como del sector privado español, a utilizar las partidas presupuestarias de que se dispone para las Tecnologías de la Información y la Comunicación simplemente adquiriendo, tanto hardware como software, licencias, permisos, en general recursos, producidos fuera de España, y que ese era justamente el camino contrario al que se debería emprender para crear una industria cultural europea (siguiendo el 'Objetivo de Lisboa') y, en particular, española, verdaderamente competente. Desde luego, no se trata de promover la adquisición de sistemas informáticos de baja calidad, únicamente por el hecho de haber sido desarrollados en España, sino de fomentar el desarrollo de sistemas informáticos de alta calidad por el procedimiento de exigir el cumplimiento de los estándares, normativas y protocolos internacionales actualmente vigentes y aplicados en países más avanzados que el nuestro en las Tecnologías de la Información y la Comunicación.

En último término, lo que aquí se plantea y en definitiva se propone, es que, a partir de las funciones que le son propias a una Biblioteca Nacional, se creen unos recursos digitales que, como se verá a lo largo de todo este texto, sean conformes a una normativa determinada, porque de ellos se seguirá una muy superior capacidad de recuperación de la información por parte, no sólo de los especialistas, sino de cualquier tipo de usuario, así como una nueva información, así mismo digital, fruto de las interrelaciones que se establecerán entre las distintas partes que componen y estructuran la información bibliográfica. Por último, se propone el desarrollo de nuevos sistemas de recuperación de la información, basados en los mismos estándares y protocolos mencionados, y que mediante procesos iterativos, configuren nuevos conjuntos de información bien ordenada.

En los últimos tiempos, el aldabonazo que ha supuesto el proyecto de Google de crear una biblioteca digital con 15 millones de obras digitalizadas, ha alarmado a algunos países europeos, que se ven forzados, o bien a responder a esa iniciativa, o bien a permitir que todos los recursos digitales referentes a Europa, acaben siendo consultados en bases de datos creadas en los E.E.U.U.

Es evidente que no se trata sencillamente de apostar por una industria europea de la Tecnología de la Información y las Comunicaciones de inferior calidad que la norteamericana y adquirirla o implementarla, por el mero hecho de haberse desarrollado en Europa. No, se trata de realizar productos de igual o

[2] *La información bibliográfica y las telecomunicaciones : estado de la cuestión y perspectivas futuras* // Xavier Agenjo: *Tratado básico de biblioteconomía* / José Antonio Magán Wals (coordinador) . - Madrid : Editorial Complutense, 1995.- ISBN: 84-7491-750-6, pp. 65-82.
Hay varias reimpresiones posteriores sin correcciones.

superior calidad, perfectamente competitivos, y que acaben evitando esta dependencia latente (por el momento) de los grandes centros de información digital norteamericanos. Hay que decir que esta situación dista mucho de ser nueva. Ya los antiguos *terminalistas*, solían fundamentalmente conectarse a DIALOG, en Palo Alto, California, de forma prácticamente exclusiva, con algunas excepciones francesas o inglesas.

El mundo de aquellas bases de datos, la tendencia a recopilar la información y a ponerla (pagando) a disposición de todo el mundo desde los E.E.U.U., no se ha modificado; antes bien, se ha intensificado, pues aquella tendencia no ha hecho más que producir en los tiempos modernos un dominio mayor. Podría decirse, en cierto sentido que los sistemas de información de los Estados Unidos, considerados en el más amplio sentido de la expresión, recogen lo que llevan sembrando durante décadas y décadas.

Aquella realidad se ha transformado en un predominio, ya en el mundo TCP/IP, es decir, el *Internet Protocol*, de aplicaciones verticales desarrolladas para archivos, bibliotecas y museos, utilizando las dos aplicaciones horizontales más importantes del mundo Internet: el correo SMTP y la red de redes, la *World Wide Web*.

Si en algún caso no queda ninguna duda de que la biblioteca puede y debe ser un productor de recursos digitales, es en el caso de las Bibliotecas Nacionales, pues estas se crean en función de una serie de principios, acerca de los cuales se ha vertido una abundatísima bibliografía[3] y, aunque parezca imposible, se ha dictado una todavía más abundante legislación, pero que en el fondo, podría sintetizarse en cuatro puntos:

a) La conservación y difusión del patrimonio documental.
b) La formación de la colección Bibliográfica Nacional.
c) El acceso a la información y al documento.
d) La coordinación y el fomento de la normativa técnica biblioteconómica en el más amplio sentido de la palabra.

Es evidente que la Biblioteca Valenciana, que tiene asignada por Ley, quizá con distintos enunciados, esas funciones puede, quizá deba, dar los pasos necesarios para que los distintos procesos biblioteconómicos que están enumerando en los cuatro puntos anteriores, puedan transformarse –incluso el cuarto- en recursos digitales, estructurándolos, eso sí, de forma que, aún partiendo de la misma base formativa, la recuperación del patrimonio bibliográfico, el Catálogo Colectivo del Patrimonio Bibliográfico, el Control de

[3] Entre las numerosas publicaciones que existen, me voy a permitir recomendar aquí el libro *Las bibliotecas nacionales : un estado de la cuestión* / Juan José Fuentes Romero . - Gijón : Trea, [2003], no sólo por la información que recopila, sino, porque el autor presenta un criterio propio y original a la hora de analizar la documentación que aporta, muy bien escogida además.

Depósito Legal, el establecimiento de la Colección Valenciana, etc., no sólo se beneficie, sino que sea la base para la creación de unos recursos digitales *específicos* que nadie más, porque nadie dispone de esa materia prima que es el Depósito Legal, o el Catálogo Colectivo, o las Colecciones Históricas, etc., puede poner en valor mediante la creación de un conjunto de técnicas, estándares y, lo que podríamos denominar soluciones informáticas, verdaderamente ricas y al mismo tiempo precisas, suministradoras de nuevos recursos digitales[4].

Al final de este trabajo se entrará en el detalle de las estructuras de información concretas y las funcionalidades específicas que en último término, han de solventar este nuevo tipo de aplicaciones sobre los que va a versar, de forma detallada, este trabajo, y que, como se va a insistir reiteradamente, tiene como objeto la creación de soluciones informáticas o *pieces of software*, que no sólo sean capaces de solucionar eficazmente, optimizando los procesos biblioteconómicos tradicionales, sino de añadir nuevas funcionalidades basadas fundamentalmente en el acceso y la difusión inteligente de la información, de modo tal que sea posible crear (y no adquirir) nuevos recursos electrónicos que estén al servicio de los usuarios de la biblioteca, así como de la biblioteca misma, a la hora de elaborar sus herramientas bibliográficas por procedimientos avanzados.

2. Principales problemas y retos

Se da por supuesto que los cuatro puntos, citados en el epígrafe anterior, están sujetos al proceso normal de automatización y digitalización que se está llevando a cabo en España (de forma un tanto desorganizada), aunque sin duda, el proyecto que promueve el Ministerio de Cultura para crear un 'directorio y recolector de recursos digitales' facilitará considerablemente la tarea para permitir que sea factible la visibilidad de los recursos digitales; los problemas de localización de este tipo de recursos, la mejora de su búsqueda y recuperación y, adelantándonos un poco a lo que se dirá más adelante, a las adaptaciones a los métodos de funcionamiento de la web semántica.

Ahora bien, los recursos digitales presentan limitaciones importantes, desde el punto de vista funcional de los buscadores a la hora de conseguir una plena visibilidad. Hasta hace poco tiempo, se limitaban a páginas estáticas HTML, aunque Google, por ejemplo, ha incorporado documentos en PDF, Word, y otro tipo de archivos; sin embargo, sólo buscan hasta un determinado nivel jerárquico, y por ello resultan inaccesibles las bases de datos, es decir, la

[4] En el sentido señalado al principio, que no debe rechazarse, excepto en el caso de que se aplique con exclusividad, resulta modelo el trabajo *La biblioteca digital de Cataluña: oportunidades, opciones y estrategias en la adquisición compartida de información electrónica* / Lluis Anglada y Nuria Comellas // *JBIDI 2000 : Primeras Jornadas de Bibliotecas Digitales, 6 y 7 de noviembre, Valladolid, 2000*. - ISBN 84-8448-066-6. - pp. 237-248. [El texto completo del artículo está disponible, gracias a DIALNET, en mi opinión, la mejor de las bases de datos documentales actualmente en España, en http://imhotep.unizar.es/jbidi/jbidi2000/22_2000.pdf].

denominada 'deep web'; no digamos, pues, si se trata de recursos digitales organizados en bases de datos. La solución que está fomentando la Comunidad Internacional, se basa en el uso de la recolección de metadatos[5] (*metadata harvesting*) sobre lo que existe en España cierta literatura que, sin embargo, pone de manifiesto una escasa familiaridad de los autores con el desarrollo real, con la creación de programas concretos de estas características, así como de soluciones informáticas que abarquen esa realidad.[6]

En resumen, las bibliotecas que reúnen información, no se puede decir aún que la recolecten, de una forma laboriosa, pero, con frecuencia inaccesible para el usuario final, padecen de una carencia fundamental, y es que, en general, disponen de herramientas para buscar, pero están aún muy lejos de disponer de herramientas que les permitan ser consultados eficazmente por terceros. No se trata tanto de ofrecer, mediante los catálogos, la información que se tiene reunida, sino de crear y desarrollar herramientas mas potentes que permitan ofrecer toda la información, aunque ésta no se encuentre, en forma catalográfica legible por el ordenador. En último término, estas fichas en formato MARC no dejan de ser sino documentos secundarios que, en poco, facilitan la visibilidad de los contenidos.

3. Visibilidad de los recursos digitales

La primera tarea que ineludiblemente han de emprender las bibliotecas que deseen un máximo de visibilidad de sus recursos digitales, consiste en la creación de un nuevo tipo de estructura de información, que constituye la materia prima, como se verá más adelante, de estos nuevos dispositivos. Me refiero a los metadatos. Gracias a ellos, páginas web, documentos HTML, bases de datos... pueden ser recolectados conforme a lo que se ha indicado en el artículo citado a pie de página anteriormente, de una forma verdaderamente eficaz. Son ya millones los recursos digitales accesibles en la red mediante el protocolo OAI en los distintos *harvester* que existen. 'OAIster'[7], al haber sido el programa creado por la Universidad de Michigan, impulsora del protocolo

[5] *Border Crossings Reflections on a Decade of Metadata Consensus Building* / Stuart L. Weibel // *D-Lib Magazine* July/August 2005. - ISSN 1082-9873. – v. 11, n. 7/8.
http://www.dlib.org/dlib/july05/weibel/07weibel.html.

[6] Tan recientemente que en este artículo, leído en abril de 2005, y redactado en el otoño de 2004, se hacían referencias a esas limitaciones: *La recolección de metadatos (metadata harvesting) y su aplicación en España* / Xavier Agenjo Bullón y Francisca Hernández Carrascal // *9 Jornadas Españolas de Documentación*, Madrid, 14 y 15 de abril de 2005, Palacio Municipal de Congresos ; [organiza, FESABID ; coordina, SEDIC]. - [Madrid] : FESABID, 2005. - 586 p. - ISBN 84-930335-5-3
Además del texto de la comunicación, puede verse la presentación, que tuvo lugar el viernes, 19 de abril, http://www.fesabid.org/madrid2005/descargas/presentaciones/comunicaciones/hernandez_francisca.pps.

[7] http://oaister.umdl.umich.edu/o/oaister . Es conveniente visitar con frecuencia este sitio, pues el incremento constante del número de repositorios, a los que accede el recolector, permite o puede permitir que una búsqueda que no había obtenido éxito en una consulta anterior, lo consiga en una segunda sesión.

OAI[8], es uno de los más populares, pero 'Scirus'[9], etc., son también increíblemente potentes. En el momento de dar la última redacción a ese trabajo, 22 de julio de 2005, OAIster reunía 5,704,392 registros procedentes de 510 instituciones.

Por supuesto, todo ello nos lleva a la necesidad de mantener una política sistemática de asignación de metadatos, es decir, de los metadatos establecidos en el DCMI ISO 15.836 para que puedan ser recuperables. Es indiscutible que sólo mediante una adopción sistemática de una política de catalogación cooperativa o federada, que reduzca al mínimo el esfuerzo en la creación de nuevos registros bibliográficos, será posible destinar los recursos sobrantes u optimizados por esos procesos, tanto para la creación de metadatos como, y esto es un proyecto realmente importante, lo que saldrá brevemente más adelante, del inicio de lo que podríamos llamar la 'subsección española del VIAF'[10] *(Virtual International Authority File*).

Naturalmente, la adopción y asignación de metadatos para los registros bibliográficos, no tendrá un sentido último, si no se lleva a cabo un esfuerzo en el desarrollo de programas que permitan la recolección, tanto en la web, como en los propios metadatos, utilizando los estándares actualmente vigentes y en pleno desarrollo. Sin duda alguna, el estándar más importante en este ámbito, es el OAI (*Open Archives Initiative*), al que se ha hecho referencia más arriba, y del que se proporciona una información detallada en la ponencia ya citada. Quizá convenga únicamente señalar la importantísima sinergia que ha surgido entre OAI y los protocolos de recuperación de información, dentro de ZING (*Z39.50 International Next Generation*) y denominados SRW[11] y SRU. En el primer caso, se denomina así porque el protocolo sobre el que resulta operativo, es a través de URL, y en el segundo caso, a través de SOAP.

---

[8] http://www.openarchives.org/OAI/openarchivesprotocol.html. En general, toda la información que presenta la página principal, se caracteriza por estar perfectamente ordenada y sea sumamente inteligible. Es una página muy viva que presenta con frecuencia nuevas funcionalidades o versiones más actualizadas de los distintos protocolos y estructuras de información que contiene.

[9] http://www.scirus.com/srsapp . Scirus nació con una vocación específicamente científica, y aunque no recoge, al menos en la actualidad, tantos repositorios como OAIster, sí plantea una filosofía y una metodología completamente distinta, que es muy recomendable conocer.

[10] Por su accesibilidad, me permito remitir al artículo que escribí con Francisca Hernández, titulado *Influencia del ICABS en el futuro digital de las bibliotecas*, y que se puede consultar en la red, en la siguiente dirección: http://www.anabad.org/admin/archivo/docdow.php?id=121.

Del proyecto VIAF, se habla en la página 6 y más adelante, en la 9. Los artículos de Bárbara Tillet sobre este programa son, sin duda, la mejor fuente de información acerca del estado en el que se encuentra y su conexión con otros proyectos y desarrollos de la IFLA y otros organismos internacionales. En cualquier buscador, pueden encontrarse diversos trabajos de la Sra. Tillet, entre los que destacaría *Authority Control: State of the Art and New Perspectives:*http://eprints.rclis.org/archive/00000332/01/tillett_eng.pdf

[11] http://www.loc.gov/z3950/agency/zing/srw

La sinergia surgida entre OAI y SRW/U[12] era quizá previsible, puesto que en último término, ambos protocolos no dejan de buscar lo mismo, es decir, el acceso de la difusión de la información bibliográfica, a partir de determinados estándares que en último término se basan en aquella remota Z39.2, origen de tantas cosas.

En último término, parecería como que existiera un plan estratégico a escala internacional, y quizá lo hay, aunque de forma no explícita, que persigue unos fines que son inherentes a la biblioteconomía. Lo que sorprende a un espectador español de todo este proceso, es la facilidad con que los profesionales de un número muy elevado de países desarrollados, logran ponerse de acuerdo, y ello desde hace ya muchos años, tanto en sus asociaciones profesionales nacionales, como internacionales, tanto en los organismos de normalización nacionales, como en los organismos de normalización internacionales, cooperando, de forma sistemática, conforme a unos protocolos de actuación perfectamente establecidos, y que sin duda, deben mucho a la teoría de sistemas y, en general y por mencionar un título clásico, a la lógica del descubrimiento científico.

4. El modelo OAI PMH

El protocolo OAI–PMH (*Open Archives Initiative-Protocol for Metadata Harvesting*) ha surgido en el entorno de la comunidad académica y científica, encaminado fundamentalmente, para la búsqueda y recuperación de textos electrónicos, supone, desde luego, una visión nueva del protocolo de búsqueda y recuperación, y se aleja del modelo de búsquedas distribuidas, como Z39.50, debido, sobre todo, a la complejidad de la aplicación de este protocolo y a la falta de precisión –tan bien conocida por los usuarios habituales de este procedimiento, cuando consultan más de un servidor a la vez- a causa de los diferentes grados de aplicación de la norma. Sin embargo, hay que decir, conforme a lo señalado anteriormente, que los nuevos desarrollos de Z39.50, hasta llegar al actual ZING, escasísimamente representado en España, se están aproximando, como con frecuencia ocurre con la normativa verdaderamente eficaz, al protocolo OAI, justamente a través de SRW/U. También se ha señalado anteriormente que la recolección de metadatos siguiendo este protocolo, forma parte de la acción conjunta, ya citada, de *IFLA-CDNL Alliance or Bibliographic Standards* (ICABS), y que tiene como objetivo promover métodos para conseguir publicaciones en la red mediante la recolección de metadatos. A la sinergia ya citada entre OAI-PMH y ZING (SRW/U) hay que añadir *Open URL*.

[12] Es muy interesante la lectura del artículo aparecido en febrero de 2005, en D-lib Magazine, *SRW/U with OAI: Expected and Unexpected Synergies* / Robert Sanderson, Jeffrey Young, Ralph LeVan, que puede consultarse en línea en http://www.dlib.org/dlib/february05/sanderson/02sanderson.html.

También se pretende aplicar el modelo OAI para la conservación a largo plazo de los recursos electrónicos, aspecto que no se aborda en esta ponencia, por ser justamente el tema de la que ha redactado Luis Ángel García Melero para estas mismas jornadas valencianas. Por último, su importancia en el desarrollo del VIAF, absolutamente imprescindible para una búsqueda verdaderamente global, ya ha quedado señalado en el trabajo de Bárbara Tillet, citado en la nota 8. Sin duda, en el próximo Congreso de la IFLA, en Oslo, se darán a conocer los desarrollos y nuevas experiencias que habrán tenido lugar a lo largo de este curso biblioteconómico mundial. No debe olvidarse que el plan estratégico del ICABS esta justamente fijado para el período 2004-2005[13].

Los componentes del modelo OAI están basados en un recolector de metadatos o *harvester*, en una interfaz de búsqueda y recuperación (que debería seguir los requisitos funcionales establecidos en FRBR[14], aunque para ello tanto las descripciones en MARC como en *Dublin Core* necesitarían de una fuerte adaptación) y un repositorio común formado por la recolección de repositorios individuales. Las peticiones y respuestas se realizan a través de HTTP.

Es importante señalar que las especificaciones para definir la estructura que permite diseñar, tanto el recolector[15] como los repositorios[16], pueden encontrarse en la dirección del proyecto *Open Archive Initiative*, ya citado más arriba, y en concreto, en las direcciones que se citan en nota. Me importa señalar, desde una perspectiva bibliotecaria que, aunque lógicamente, es necesaria la figura de un analista programador para llevar a cabo el desarrollo y poner en funcionamiento estas piezas de software, que su estructura es absolutamente inteligible para un archivero, para un bibliotecario, o para un museólogo. La primera impresión, como ocurre ante cualquier lenguaje de marcado, parece presuponer una codificación difícil, pero no lo es mucho más de lo que pueda serlo el formato MARC, y por lo tanto, debe ser asumida por los profesionales de las instituciones de memoria y muy en particular por los bibliotecarios, e impulsar el desarrollo de los sistemas de recuperación basados en este modelo y en el citado protocolo.

Las características básicas que pretende fundamentalmente este protocolo son, por un lado, una flexibilidad para adaptarse a cualquier ámbito, y por otro lado, una gran facilidad de implantación. Ofrecen información sobre

---

[13] http://www.ifla.org/VI/7/annual/icabs-sp2004-2005.pdf

[14] Existe traducción española de Xavier Agenjo y María Luisa Martínez-Conde, impresa en papel y accesible en la red: http://travesia.mcu.es/documentos/requisitos.pdf .

[15] http://www.openarchives.org/OAI/openarchivesprotocol.html. Véase lo dicho en la nota 6.

[16] http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm. La creación de repositorios estáticos, conviene insistir, no supone necesariamente el desarrollo de un *harvester*, sino que es sencillamente una forma de roturar el campo para que terceros recolectores recuperen la información, gracias a ese repositorio.

cualquier tipo de recurso, tanto físico como digital, debido justamente al uso de los metadatos.

Presentan una gran flexibilidad en la formación de repositorios, siempre y cuando se ajusten a la estructura ya citada, y desde luego, soportan tanto bases de datos, como meros ficheros estructurados. También transmiten y presentan la información en varios formatos, y por supuesto, son los básicos para los profesionales que he citado anteriormente, la EAD, el XML MARC y el MLA *Spectrum*, así como los más desestructurados *ePrints*. El formato mínimo es el *Dublin Core* sin cualificar, todo lo cual permite un serie de mecanismos para la recuperación total o selectiva. Quizá convenga señalar aquí de nuevo el ya citado artículo, presentado al 9 Congreso de Fesabid, al que habría que añadir el muy interesante proyecto MINERVA[17] *(Mapping the Internet Electronic Resources Virtual Archive*), o el *WayBack Machine*[18], o citar los acuerdos, tanto de Yahoo, con de la Universidad de Michigan, como el de Google y la Biblioteca Nacional de Australia, de marzo de 2005. No debe olvidarse en ningún caso, que el proyecto *mod_oai*, financiado por la *Andrew Mellow Foundation*, pretende el desarrollo del software para transformar la información contenida en servidores Apache, y transformarlos en servidores OAI y, lo que es importantísimo, que el *mod_oai* se desarrolla bajo licencia pública (GNU *Public Licence*).

En definitiva, lo que aquí se propone es que la creación de los registros bibliográficos y los registros digitales asociados en su caso a los primeros, incorporen siempre, como mínimo, un conjunto de metadatos estructurados según DCMI-ISO 15834, pero que no se detenga ahí la iniciativa, sino que simultáneamente se desarrolle un repositorio estático en el que descargar sistemáticamente los registros, de forma tal, que puedan ser recolectados por *harvesters*, desarrollados por terceros; dar de alta nuestros repositorios en los más importantes *harvesters* existentes, como los ya citados más arriba, y por último, y como instrumento fundamental para completar nuestra colección de recursos digitales sobre la materia que le es específica, y que puede a su vez recolectar en la red recursos digitales con que incrementar nuestra biblioteca digital[19], verdadera biblioteca virtual[20] de recursos generados, tanto mediante la

[17] http://www.loc.gov/minerva/collect/elec2002

[18] http://www.waybackmachine.org

[19] En el muy interesante número de julio-agosto, vol. XI nº 7/8 (2005) del D-lib Magazine, se pueden consultar dos artículos sobre el presente y el futuro de estas iniciativas. *Digital Libraries Challenges and Influential Work* / William H. Mischo: http://www.dlib.org/dlib/july05/mischo/07mischo.html, así como *Where Do We Go From Here? The Next Decade for Digital Libraries* / Clifford Lynch: http://www.dlib.org/dlib/july05/lynch/07lynch.html.

[20] Sobre la falta de decantación entre las expresiones 'biblioteca digital' y 'biblioteca virtual', debe leerse el artículo que se cita a continuación. Aún reconociendo su rigor expositivo, debo decir que yo no estoy de acuerdo con sus conclusiones: *Propuestas de concepto y definición de la biblioteca digital* / Jesús Tramullas Saz // *III Jornadas de Bibliotecas Digitales : (JBIDI'02)* : El Escorial (Madrid) 18-19 de

digitalización de los fondos propios, como de la recolección de fondos producidos por terceros sobre nuestra propia materia. Así, en una consulta realizada el 19 de julio de 2005, sobre las distintas *ePrints* existentes en España, se obtiene un total de 27 recursos digitales de tesis doctorales (se entiende que con su dirección URL completa) defendidas en bibliotecas españolas, así como algunos artículos de investigación en revistas especializadas, sobre el descriptor 'Valencia'

5. La web semántica

Son tantas las definiciones que existen sobre esta nueva concepción de la red, que quizá lo más práctico sea remitir al interesado a la página web del W3C[21] y permitir que allí se beba en las fuentes directas[22]. En último término, la web semántica es un nuevo modelo de estructuración, interrelación y recuperación de la información, gracias al uso de una serie de normas y estándares ya existentes, como XML o RDF, o algunas nuevas creadas al efecto, como OWL[23], así como determinados protocolos de comunicación y estructura de la información, que configuran una solución informática, completamente nueva y original, y sumamente potente. Pretende superar las carencias e insuficiencias que la *World Wide Web* histórica ha ido presentando de forma directa al crecimiento exponencial de número de sitios en la red. En definitiva, lo que se pretende es sustituir los sitios tradicionales por unas nuevas estructuras de información publicadas en la red, a las que se denomina 'ontologías', el conjunto de las cuales constituirán, junto con los agentes de software intermediarios, la web semántica, así como el conjunto de los sitios actuales constituyen en la actualidad la *World Wide Web* que conocemos[24].

Noviembre de 2002 / coord. por José Hilario Canós Cerdá, Purificación García Delgado, 2002, ISBN 84-688-0205-0, pp. 11-20: http://mariachi.dsic.upv.es/jbidi/jbidi2002/Camera-ready/Sesion1/S1-1.pdf.

[21] http:// www.w3c.es. Esta dirección es la versión española del www.w3c.org, lo que siempre facilita la consulta, aunque con frecuencia haya que acudir a la fuente original.

[22] Aunque cerrado el año pasado (2004), contiene el más amplio conjunto de menciones y declaraciones sobre la ontología. Lo considero de lectura imprescindible. http://www.w3.org/2001/sw/WebOnt

[23] http://www.w3c.es/Traducciones/es/SW/2005/owlfaq. Esta página, muy bien traducida al español, no sólo es un buen conjunto de explicaciones sobre la OWL, sino que también aclara algunos puntos sobre el concepto y la función de la Ontología.

[24] Sobre la bibliografía en español, acerca de la ontología y la web semántica, me remito a nuestro artículo, publicado en el II Congreso de Bibliotecas Públicas Españolas, *De la Biblioteca Virtual a la Ontología y la Web Semántica* / Xavier Agenjo, Francisca Hernández. Además de la edición en papel y en disco, puede consultarse en Internet en la dirección del Ministerio de Cultura, y dentro de la sección Travesía: http://travesia.mcu.es/documentos/congreso_2bp/3a_sesion/comunicacion01.pdf

La *World Wide Web* que conocemos, puede caracterizarse, entre otras muchas maneras, por estas cinco características:

1) Los recursos están relacionados únicamente mediante enlaces simples que denominamos URLs.

2) Los usuarios de la *World Wide Web* navegan entre los referidos recursos gracias a los citados enlaces.

3) Ahora bien, los sistemas informáticos en la web actual, no disponen de otra información significativa que no sea la dirección del recurso.

4) Por otro lado, el contenido de la página a la que se ha navegado, sólo es inteligible por el usuario.

5) Por último, los sistemas informáticos sólo son capaces de leer la dirección, pero no pueden interpretar el contenido.

Cuando surgió la necesidad, y por lo tanto el motivo, de hacer un nuevo diseño de la web, se partió del hecho de que se calculaba en que 1996, el 50% de los usuarios encontraban información en la web sin necesidad de llevar a cabo unas búsquedas complejas. Sin embargo, en el año 2002, el 40% de los usuarios terminaban las sesiones de búsqueda sin los resultados apetecidos. Quizá por ello, en esa fecha es cuando se redacta el artículo, para decirlo según la expresión inglesa, seminal, sobre la web semántica[25], artículo sobre la segunda WWW escrito por quien también creó la primera.

Las nuevas funcionalidades, tanto de Google como de Yahoo, u otros buscadores muy potentes, no logran, sin embargo, superar las limitaciones de la recuperación de la información y, desde luego, se presentan dos extremos contrapuestos, que desorientan, cuando no imposibilitan completamente, el trabajo del usuario. Por un lado, información oculta, como podrían ser las bases de datos, y por otro lado, la frecuentísima presencia del exceso de información, el denominado 'ruido'. De hecho, y tal y como previó en su día Claude Shannon (1916-2001)[26] en su *Teoría matemática de la comunicación* (1948), los ruidos y silencios son inherentes a los sistemas de información automatizada.

A todo ello hay que añadir, y esto es fundamental, que existe una inmensa cantidad de información todavía no digitalizada. Quizá uno de los aspectos, no ya tácticos, sino estratégicos, del proceso global de digitalización, consiste en

[25] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). *The Semantic Web*. // Scientific American . 279(5). El texto está disponible en: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2. [El artículo estaba citado, según *Google Scholar*, en 1542 trabajos, el 21.07.05]

[26] Una información completa, entre las aproximadamente 64.500 páginas dedicadas a 'Claude Shannon', figura básica de nuestro tiempo, es http://www-groups.dcs.st-and.ac.uk/%7ehistory/Mathematicians/Shannon.html.

llevarla a cabo, no sólo para funciones de preservación o incluso de consulta simple, sino para edificar en torno a ellas, y basada en ellas, estructuras complejas de información como podrían ser los repositorios y los recolectores basados en DCMI o en OAI, como se ha dicho en la primera parte de este trabajo, o más aún, edificando ontologías que permitan construir a medio plazo, una constelación de sitios, conforme a la metodología de la normativa definida por el W3C, y que se acaba de referenciar, y que tenga como objeto ir construyendo la futura web semántica, que será fundamentalmente un conjunto de sitios especializados, es decir, ontologías, que abarque de forma interrelacionada cualquier aspecto de la información y, sobre todo, del conocimiento.

6.Características de la web semántica

Dado que el nacimiento de las ontologías y, por lo tanto, de la web semántica, es tan reciente, son muchas las interpretaciones y diseños que actualmente se están discutiendo. Pero si vamos estableciendo una contraposición entre la web actual, tal y como la hemos enumerado más arriba, se podría decir que la estructura de la información presenta estas características:

1) Los recursos no están relacionados únicamente mediante enlaces simples, sino que tienen un significado, por ejemplo, Vicente Blasco Ibáñez *es autor* de 'La Barraca' o 'La Barraca' *es una obra* de Vicente Blasco Ibáñez.

2) Los usuarios de la *World Wide Web*, pueden ser en el caso de la web semántica, agentes de software, y gracias a este nuevo tipo de enlace cualificado, pueden interpretar las relaciones y navegar a significados más amplios o más restringidos.

3) A diferencia del modelo anterior, el contenido de la página, previamente estructurado, en la que se ha navegado es inteligible, no sólo por el usuario, sino por el programa de consulta o los agentes de software.

4) No sólo los usuarios son capaces de navegar gracias a los citados enlaces simples entre los recursos electrónicos, sino que el software asociado ayuda al usuario a recuperar una información verdaderamente pertinente.

5) Por último, los sistemas informáticos son capaces de leer la dirección y, utilizando relaciones extensibles y basadas en las definiciones de las ontologías, proporcionar un conjunto de respuestas, incluso no previstas por el usuario o por el sistema informático que se está utilizando.

7. La construcción de la web semántica.

Ya tomada la decisión de estructurar la información referida a un dominio, como podría ser en este caso la Comunidad Valenciana y su Patrimonio Bibliográfico, o los recursos topobibliográficos, es necesario dar una serie de pasos para construir la web semántica:

A) Necesidad de construir un substrato de relaciones semánticas. Por ejemplo, el Catálogo Colectivo del Patrimonio Bibliográfico puede construirse en una Biblioteca Virtual, en la cual, los conjuntos de información complementarios, listas de lugares de impresión, nombres y circunstancias de impresores, libreros y editores, las letrerías utilizadas por ellos, las bibliotecas en las que se encuentran depositados, etc., se interrelacionan conforme a un modelo de entidad/relación[27].

B) Es así mismo necesario la organización del saber de un determinado dominio, estableciendo clases, jerarquías entre clases, atributos. Es evidente que, dado el propósito de esa organización del conocimiento, el modelo que proporcionan las *FRBR*[28], es, con mucho, el más apropiado para el entorno bibliográfico, aunque habría que tener más que presente el *Conceptual Reference Model*[29]. Este modelo, facilita enormemente la tarea de construcción de ontologías, no sólo por su riqueza conceptual y estructural, sino porque constituye un marco de referencia para la totalidad del patrimonio cultural. En el marco de una ontología para el patrimonio bibliográfico difícilmente podrían cortarse las estrechísimas relaciones entre los diferentes patrimonios culturales por lo que un modelo conceptual más amplio es absolutamente necesario.

---

[27] *Del Catálogo Colectivo a la Biblioteca Virtual : La Biblioteca Virtual del Patrimonio Bibliográfico* / Xavier Agenjo y Francisca Hernández //*I Jornadas sobre Patrimonio Bibliográfico en Castilla-La Mancha*, celebradas en Toledo los días 12, 13 y 14 de noviembre de 2003 y organizadas por el Servicio Regional del Libro, Archivos y Bibliotecas de esa Comunidad

[28] Sobre el estado de la cuestión de la armonización de FRBR y CRM, puede verse esta presentación: http://www.oclc.org/research/events/frbrworkshop/presentations/zumer/Manifestation_and_attributes.ppt presentada al Congreso auspiciado por el OCLC, donde numerosas presentaciones y trabajos son de obligada lectura. http://www.oclc.org/research/events/frbr-workshop/

[29] http://cidoc.ics.forth.gr/official_release_cidoc.html. En esta misma página se ofrece una codificación del modelo (versión 3.4.9) en RDFS, directamente importable a programas de edición de ontologías como Protégé [http://protege.stanford.edu/].

La versión 3.4.9 se corresponde, desde octubre de 2003, con la ISO Draft International Standard submission ISO/DIS 21127.

Desde el 5 de julio, la versión 3.1 de este extendido editor incluye un potente y útil editor de OWL [http://protege.stanford.edu/plugins/owl/]

.

C) Como ya se ha dicho, la web semántica será la suma coordinada de ontologías de dominios concretos, el Catálogo Colectivo del Patrimonio Bibliográfico valenciano, por ejemplo, más el Depósito Legal de Valencia, más la Colección Valenciana, más la Colección de Referencia, constantemente incrementada mediante recolecciones de metadatos, constituirían una ontología o un conjunto de ontologías, que formarían parte de una web semántica de carácter bibliográfico-cultural, y en general, vinculado con las instituciones de la memoria.

D) Es muy importante tener presente que, gran parte de la eficacia de la consulta a una ontología, es que el usuario se beneficia de la navegación a través de las relaciones definidas en esas ontologías.

Con frecuencia, aquellos profesionales que se sienten interesados por las perspectivas que parecen ofrecer las ontologías para las instituciones de memoria, lamentan no encontrar entre las muchas ontologías que se están desarrollando en los últimos años, ninguna que encaje exactamente con su campo de competencia. Se podría mencionar, tal vez, por su interés en un entorno europeo, dos iniciativas específicas: EuroStory.net[30] y VICODI[31], que sí participan, y de forma avanzada, de este nuevo modelo conceptual e informático.

8. Objetivos finales

Los objetivos finales pueden resumirse en cuatro puntos:

1) Se pretende proporcionar diferentes tipos de información a diferentes tipos de usuarios, investigadores, especialistas, docentes, usuarios de carácter general, puesto que justamente algunos de los atributos definidos en las relaciones así lo permiten.

2) Crear servicios y contenidos para la Educación. Actualmente, y dada la enseñanza virtual y las aulas virtuales, están proliferando extraordinariamente los *recursos didácticos*[32] de carácter digital. Es evidente

[30] http://www.eurohistory.net/Index.do.

[31] http://www.vicodi.org/about.htm. Muy clarificadora resulta, en este sentido, la siguiente presentación: http://www.museumscomputergroup.org.uk/meetings/1_2005_docs/A-Beginner's-guide-to-the-Semantic-Web.ppt

[32] Debe leerse, por su interés y novedad, *Desarrollo de repositorios de objetos de aprendizaje a través de la reutilización de los metadatos de una colección digital : de Dublin Core a IMS* / Clara López Guzmán, Francisco José García Peñalvo, Pedro Pernías Peco // *RED: Revista de Educación a Distancia*, ISSN 1578-7680, Nº. 2, 2005 (Ejemplar dedicado a: I Simposio Pluridisciplinar sobre Diseño, Evaluación y Descripción de Contenidos Educativos Reutilizables): http://www.um.es/ead/red/M2/lopez27.pdf.

que las instituciones de memoria, a través de ontologías, o bien directamente, o bien a través de una pasarela, pueden suministrar contenidos digitales que puedan tener valor para esas colecciones de recursos didácticos, de forma verdaderamente considerable. Un aspecto central sería, por ejemplo, la de recursos didácticos de carácter digital para los portales ELE, es decir, la enseñanza del español como lengua extranjera en un entorno automatizado y digital.

3) Facilitar un acceso completo a esos recursos digitales mediante exposiciones virtuales, los ya citados recursos didácticos o la creación de bibliotecas digitales mediante el 'escaneo' de los documentos primarios, estructurados en forma de colecciones[33]. Es muy importante tener en cuenta que la Unión Europea acaba de lanzar el programa *i2010 (A European Information Society for growth and employment*)[34], entre cuyos cuatro objetivos fundamentales está justamente el de la creación de bibliotecas digitales[35].

4) Gracias a una ontología se podrá mostrar todos y cada uno de sus objetos, así como sus relaciones, a través de metadatos y de las vinculaciones con otro tipo de objetos que posean categorías y atributos similares.

Conviene decir que las denominadas instituciones de memoria, es decir archivos, bibliotecas y museos, están en una posición inmejorable para llegar a conseguir todo el potencial de la web semántica. En efecto, hay que tener en cuenta, que estas instituciones estructuran la información que poseen en instrumentos tales como inventarios, catálogos, tesauros, clasificaciones y todo tipo de taxonomías. Una gran mayoría de esos instrumentos se encuentra ya automatizada[36], y un porcentaje elevado en estructuras normalizadas. Ya son menos las que han utilizado lenguajes de marcado para estructurar esta información, y menos aún las que emplean metadatos para atribuir significados mediante una codificación precisa. Sin embargo, los dos últimos pasos no presentan dificultades especiales, y así, pasar del 'etiquetado' MARC al XML MARC Schema y/o al DCMI, no entraña una especial problemática, sencillamente, los responsables de esa institución de memoria tienen que conocer y, por lo tanto, darse cuenta y tomar la oportuna decisión para realizar

[33] Aunque probablemente muy superado por el 'Directorio y recolector de recursos digitales OAI' que está preparando el Ministerio de Cultura, es útil la consulta del trabajo '*Bibliotecas digitales españolas a texto completo*' / Josep Lluis Canet // *Syntagma: revista del Instituto de Historia del Libro y la Lectura*.- ISSN: 1695-6958,1 (2005) 149-159.

[34] http://europa.eu.int/information_society/eeurope/i2010/index_en.htm

[35] Para tener una información más completa sobre este importante proyecto cuya trascendencia no puede exagerarse, puede consultarse la dirección http://europa.eu.int/i2010

[36] Así, por ejemplo, *Aplicación integrada de la Biblioteca Digital del Patrimonio Histórico Andaluz* / María José Escalona, M. Mejías, Jesús Torres, Juan M. Cordero, M.G. Romano / *JBIDI 2000 : primeras Jornadas de bibliotecas digitales*, 6 y 7 de noviembre, Valladolid, 2000, ISBN 84-8448-066-6, pp. 295-298: http://imhotep.unizar.es/jbidi/jbidi2000/31_2000.pdf.

las transformaciones de los calificadores de contenidos a las nuevas etiquetas XML o DCMI e incluso RDF.

En mi opinión, darán ese paso si perciben la importancia de cumplir las funciones biblioteconómicas que le son propias y que pueden alcanzar mediante cosechadores y ontologías, realizando las transformaciones necesarias de esos códigos a las nuevas estructuras de información.

Es claro que el desarrollo de la web semántica depende del marcado sintáctico y semántico del contenido, así como del desarrollo de herramientas que permitan el análisis del conocimiento. Para ello se deberán adoptar lenguajes de representación del conocimiento, como XML, RDF, u OWL, para lo cual, y como se acaba de decir en el párrafo anterior, se encuentran también en situación privilegiada, respecto a otras áreas de conocimiento por los argumentos ya expuestos. Más complejas son las metodologías para definir y extraer el conocimiento de un dominio, es decir, para construir las ontologías, para lo cual hace falta un conocimiento exhaustivo y fiable de dicho dominio, pero también unas herramientas informáticas que lo faciliten, al igual que los correspondientes 'mapeos'.

En la actualidad Digibis, con la colaboración de la Fundación Ignacio Larramendi, está llevando a cabo, dentro del Plan PROFIT, una iniciativa de este tipo, denominada 'Ontología y web semántica de polígrafos', FIT-350200-2004-38[37]. Confío que pueda tener una utilidad, no sólo para nuestro proyecto, sino para las bibliotecas virtuales en general, que pretendan transformarse en ontologías.

De hecho, las ya citadas tres colecciones fundamentales que pueden caracterizar el fondo específico de una biblioteca nacional, es decir, el Patrimonio Bibliográfico, la colección de la Bibliografía en curso y la 'Colección Nacional', que responden a funciones específicas de una Biblioteca Nacional, se caracterizan por suponer una recolección de información en un área o territorio determinado, en el cual la Biblioteca Nacional ejerce su acción; en coordinar la producción de recursos digitales a partir de esas colecciones bibliográficas y en transformarlas en contenidos para constituir una o varias ontologías para la web semántica.

Por citar un mero ejemplo, en torno al Patrimonio Bibliográfico, pueden recopilarse, digitalizarse, estructurarse y relacionarse con el conjunto de normas y protocolos que se han ido citando, tipo y topobibliografías, censo de impresores, libreros y editores, con las fechas de su período de trabajo, variantes de los nombres latinos, de los lugares de impresión y edición o de los propios editores y libreros, marcas de impresores, marcas de agua, repertorios

[37] Resolución de 2 de marzo de 2005, de la Dirección General para el Desarrollo de la Sociedad de la Información, por la que se publican las ayudas concedidas en el año 2004 del Programa Investigación y Desarrollo de la Sociedad de la Información.

de encuadernadores y grabadores y, en general, todo tipo de instrumento bibliográfico que habitualmente sirve de apoyo *desde fuera* al proceso de creación del registro bibliográfico, pero que ahora debería estar estructurado con él, *desde dentro*, mediante una serie de relaciones, así como con el recurso electrónico, fruto de la digitalización del mismo.

La ontología permitirá la extracción del conocimiento de esa información estructurada, estableciendo clases como la de personas, en la que habrá autores, traductores, impresores, libreros, grabadores, etc., la clase entidades, la clase lugares, la clase tiempo, etc. Lo mismo habría de hacerse con la clase obras, para lo cual, evidentemente, la metodología de las FRBR/CRM resulta una falsilla imprescindible y, por último, establecer las relaciones entre las distintas clases o entidades, las cuales nos permitirán conocer en una consulta una concatenación de respuestas en las que obtendremos las obras de un determinado autor, vertidas a otra lengua por un determinado traductor, puestas en la imprenta por un determinado impresor, que tiene su taller en un lugar preciso y en un período de tiempo determinado, pues todas esas relaciones se habrán establecido entre diferentes clases.

Por último, la secuencia finalizará con la reproducción digital del recurso electrónico y con la calidad que le puedan proporcionar los instrumentos de edición, cada día más potentes y dotados de *plug-ins* cada vez más sofisticados, que podrán desencadenarse cuando se solicite la visualización o la impresión del documento.

9. Conclusiones

Toda la exposición de recomendaciones, normativa y estructuras de información expuestos en este texto, tienen como objetivo animar a que las bibliotecas nacionales, y también el resto de instituciones de memoria, encaminen la actividad que ya tienen marcada y definida hacia el entorno de la creación de contenidos digitales y su difusión en la web por los mecanismos y tendencias más actuales de la misma. En definitiva se trataría de:

1. Hacer accesible internacionalmente el Patrimonio Cultural.

    1.1. Crear un corpus orgánico del Patrimonio Cultural: contenido digital sobre el patrimonio bibliográfico, archivístico e inmueble al que en la actualidad es difícil acceder por tratarse de obras y piezas dispersas, difíciles de localizar, o que no son accesibles al público en general.

2. Generar contenidos en los idiomas de España para la web Internet.

2.1. Digitalización de documentos, piezas, etc., conversión de bases de datos existentes, conversión de fuentes de referencia consolidadas...

3. Generar sistemas de información adaptados a la normativa de la comunidad bibliotecaria internacional y a las tendencias de las tecnologías de la información en la web.

4. Conocer, mediante el desarrollo y uso de las correspondientes aplicaciones y sistemas de información como la recolección de metadatos o la creación de ontologías, el dominio Patrimonio Cultural para poder intervenir en él.